学校编码：10384　　分类号＿＿＿＿密级＿＿＿＿

学号：X2005230030　　UDC＿＿＿＿

厦门大学

硕 士 学 位 论 文

基于 WEB 的海运信息采集与分析系统的设计与实现

Ocean Shipping Information Collection and Analysis System Based on the WEB Technology

何 昭 鹏

指导教师姓名：姜青山　教授

专 业 名 称：软 件 工 程

论文提交日期：2008 年 11 月

论文答辩时间：2008 年 12 月

学位授予日期：2008 年 12 月

答辩委员会主席：＿＿＿＿＿＿

评　　阅　　人：＿＿＿＿＿＿

2008 年 11 月

基于 WEB 的海运信息采集与分析系统的设计与实现

何昭鹏

指导教师：姜青山　教授

厦门大学

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

港口是物流运动的重要汇接点，是地区和国家间物资交换的枢纽，在对外贸易中占据了绝大的比例。港口的海运信息具有经济晴雨表的作用。从港口原料类货物运量的变动可以反映一个地区或国家的经济消长态势，这比产成品市场所反映的信息更加敏感；港口货种结构调整信息可折射一个地区或国家产业结构的变动趋势；通过对单位重量货物价值信息的深入挖掘分析，可以发现该地区或国家的经济发展现状和趋势等等。获取并分析海运信息对研究和分析经济发展有十分重要的意义。

传统的海运信息获取方法是通过定购海运管理机构和以赢利为目的的专业咨询服务机构发行的报刊或专刊来获取数据，再加以分析研究。周期长、成本高和主题固定，难以满足研究部门需要更实时、更细节和更互动的分析需求。随着世界各国港口信息化建设的不断深入，通过互联网发布船舶动态信息并提供各种海运信息咨询服务的应用日益普及，使得利用互联网来搜集海运信息成为可能。然而，由于不同港务网站信息发布的差异性和网页展现格式的多样性，造成依靠人工方式来完成信息收集几乎无法实现。

针对上述问题，本文重点分析和研究了海运信息的网页特点，设计并建立了一套基于 WEB 数据的海运信息采集与分析系统，实现了网页海运数据的自动获取、抽取转换和数据库集成并提供查询分析功能。系统包括 WEB 信息采集、ETL 处理和信息查询分析三个子系统。本课题主要研究内容如下：

1. 信息采集子系统采用 WEB 信息抽取技术完成对各种复杂网页表格数据的自动下载、解析、识别和转换工作，解决人工难以承担的难题；
2. ETL 处理子系统高效完成不同港务海运数据的转换、清洗、过滤和整合等功能，并以统一格式保存到数据库中；
3. 信息查询分析子系统提供固定格式的查询统计和报表生成功能，并通过数据仓库技术构建了联机分析处理（OLAP）应用，满足不同用户群体的多样化数据查询和分析需求；
4. 对系统进行集成和部署，并根据有关要求进行网络安全设计。

关键词： Web 信息抽取；数据仓库；ETL 工具；OLAP

Ocean Shipping Information Collection and Analysis System Based on the WEB Technology

Abstract

The harbor is an important confluent point of freight. The economical situation of a district or country can be reflected by the change of material freight gross; It has very important sense that acquiring and analyzing ocean shipping information to the research of economical development.

With the deep information-based construction of world harbors, and with the popularization of kinds of consulting service and the dynamic shipping information which be issued in the internet, It makes collecting information come true via the internet.

The thesis analyses the characteristics of the ocean shipping information, designs a system that collects and analyses ocean shipping information based on WEB data. This system realizes automatic adopting,decimating and converting ocean shipping information in web, and integrates the information in database. The system consists of three sub-system which are the WEB information adoption system,TEL disposing system,information querying and analyzing system.Our main work are as follows:

1. The information collection subsystem finishes the data downloading, parsing, recognizing, and converting work of kinds of complex web page data by adopting the technology of WEB information extraction, which solves the difficult problem of manual operation;
2. The ETL process subsystem implements effectively the data converting, cleaning, filtering, integrating function of different harbor ocean shipping, and save the formatted data into database;
3. The information inquiring subsystem provides the function of formatted inquiring statistic and building forms, and creates the OLAP application by using data warehouse technology, which satisfies different users with the diverse query and analysis;
4. System integration and deployment is accomplished. Web security design is carried out according appointed requirement.

Keywords: Web Information Extraction；Data Warehouse；ETL Tools；OLAP

目 录

Table of Contents

第一章 绪论

海上运输是地区和国家间物资交换的重要渠道，港口信息具有经济晴雨表的作用。获取并分析船舶运输信息对研究和分析经济发展有着十分重要的意义。以报刊为媒体的传统海运信息收集方法已难以满足研究部门的实时性和细节性需要，随着港口信息化建设程度不断提高，为我们通过互联网途径快速获取海运信息创造了条件。本章将对基于 WEB 的海运信息采集与分析系统的研究背景意义、应用现状以及存在问题等进行阐述，最后对本文研究内容以及本文结构安排等进行总体介绍。

1.1 研究背景及意义

港口是物流运动的重要汇接点，是地区和国家间物资交换的枢纽，通过海上运输完成的对外贸易物资占据了绝大的比例，对于那些岛屿国家来说更是如此（如日本就占 99%以上)。港口信息具有经济晴雨表的作用。港口原料类货物运量的变动可以反映一个地区或国家的经济消长态势，这比成品市场所反映的信息更加敏感；港口货种结构信息折射一个地区或国家产业结构的变动趋势；通过对单位重量货物价值信息的深入挖掘分析，可以发现该地区或国家的经济发展现状和趋势等等[1]。船舶作为港口对外运输的唯一载体，获取并分析船舶运输信息对研究和分析经济发展有着十分重要的意义[5]。

然而，传统的海运信息研究资料大多来源于国内外公开发行的刊物，必须经过收集、摘录、编辑、统计汇总后再进行分析研究，才能形成结论。如每年出版的《中国海运业研究咨询报告》，需要从国家统计局、海关总署、国际航运经济与物流研究所(ISL)等 8 个国内外机构组织，以及《海运情报》、《船舶经济贸易》、《世界海运》等 12 种公开出版物收集资料并加以分析研究[2]。其资料范围之广、工作量之大显而易见，每份报价 7000-12000 元也证明了这一点。即使如此，由于信息收集的延迟性和发行报刊本身带有综合性的特点，我们无法从中获取更实时、更细节的第一手资料，难以满足某些特定主题的研究需要[3]。为此，我们急需寻找一种新的信息获取渠道，可以更快、更好地完成资料积累。近年来，由于计算机性能提高、成本下降及数据管理技术的成功运用，港口管理部门的信息化程度越来

越高，Internet 及相关技术又使计算机、网络、通信三合为一，大大推动了世界各国对海运自动化和网络化管理进程。特别是船舶交通管制系统（VTS）[4]、甚高频通信系统（VHF）和船舶自动识别系统（AIS）等系统在世界各国和运输船舶上普及性建设，使得船舶动态信息更加及时准确。为我们通过互联网这一廉价的途径获取海运信息创造了客观条件。

1.2 研究现状及存在问题

港口信息化建设是海运信息数字化的基础，港口信息在互联网上的广泛应用是通过互联网获取海运信息的前提条件。以 Internet 为代表的现代信息技术得到越来越广泛的应用，基于 Internet /Intranet 的计算机网络系统正在替代传统的管理模式，信息的交流与共享已成为当今社会发展的主流。随着无线宽频网络应用的进一步扩展，无线通讯将成为一种廉价、可靠的通讯方式，它们在航海领域的应用已日渐成熟。用现代信息技术，特别是基于 Internet 与无线通讯技术来构筑航海信息系统将对 21 世纪的航海产生积极的影响[5]。下面简要介绍国内外港口信息化建设、信息应用情况和未来发展趋势，并对当前从互联网获取海运信息存在的困难和问题进行阐述。

1.2.1 港口信息化建设现状

随着世界经济的高速增长和全球一体化进程的加快，世界港口在全球经济发展中的突出贡献越来越明显。在面对船舶向大型化、集装箱化发展趋势下，国内外的港口企业迎来了又一个巨大的发展机遇。经济的全球化，势必带来对交通、物流的更加广泛的需求，其中，港口在全球供应链中有着不可替代的作用，过去的枢纽港已成为全球现代物流网络的重要节点。随着先进的信息技术不断应用到港口信息化建设上，促进了港口的生产经营和信息服务水平的提高。管理现代化、信息化已成为现代港口建设与发展的重要目标之一[6]。

新加坡港是全球海运中心，历来以国际转运业务为主，年处理集装箱数量占全球总量的 20%。每个月集装箱吞吐量超过 100 万 TEU。在过去 5 年新加坡港投资了一亿六千万加币在信息科技运用上，目前有超过 350 个应用系统在处理港埠管理、规划与作业上，主要有数据库查询服务，包括船舶靠港时程、集装箱/货物清单、集装箱/货物追踪及化学危险品数据库等；提供海运相关信息，船舶动态数

据；电子文件交换，集装箱舱单、危险品申报、靠港申请及出港时程预报等通关自动化，透过贸易网络(TradeNet)关贸网络网网相连，可与政府国贸及签审机关作数据交换[6]。

我国沿海主要港口信息化建设工作自二十世纪八十年代起步，历经八十年代初级使用，九十年代拥有广域网络，目前都实现了互联网络的广泛应用。近年来沿海港口结合自身的发展规划以及“两个转变”的要求，实施了以财务管理、生产管理、人力资源管理和资产管理为核心的集团资源管理系统（ERP）、决策支持系统（DSS）、办公自动化系统（OA）和客户服务等系统[6]，并逐步利用互联网提供各种在线信息服务。

在加大港口网络信息化建设的同时，世界各国也加强了对船舶监控和调度管理，为了保障近岸水域船舶航行的安全，世界各国都很重视近岸水域的安全设施建设，包括船舶交通管制系统（VTS）、甚高频通信系统（VHF）和船舶自动识别系统[7]（AIS）等。其中 AIS 系统采用自控时分多址联接技术，在海事 VHF 频段自动连续发送本船静态、动态和航次信息及安全短消息。同时也自动接收周围船舶发出的这些信息，并与海岸基站进行信息交换。国际海事组织（IMO）规定，国际航线的 300 总吨以上船舶和公约国航行于国内航线的 500 总吨以上的船舶，以及所有客船，从 2002 年 7 月 1 日起到 2008 年 7 月 1 日止分阶段装配 AIS 设备。AIS 系统对整个航运的安全和管理带来了深刻的变化。目前，国内外许多港口的港务局积极应用高精度定位（GPS）、船舶全球唯一的编码（MMSI 码）、自控时分多址联接（SOTDMA）和电子海图实时显示等技术[7]，获取船舶的船名、呼号、MMSI、船长、船宽、吃水、货物种类等船舶静态数据，生成航向、航速、位置、相对距离等船舶航行动态数据，建立了船舶动态信息预报系统，用于准确掌握在港和进出港船舶的动态，指导船舶航道疏通、靠离泊引航，加强对进出港、靠泊码头、锚地的船舶及对装卸作业进行安全监督、管理等，大大提升了港口船舶的管理能力和动态调度能力。另外，有些船舶代理机构也利用 AIS 构建了船舶代理业务监控系统[7]，加强了船舶调度的科学性，提升了船代服务质量。

目前，世界各大港口通过引进先进技术和设备，如 EDI、船舶交通服务系统（VTS）以及堆场智能化管理技术等，不断提高其管理水平和运作效率，港口业务逐步向专业化、规范化、标准化迈进。可以预见，随着全球经济一体化与信息技术的发展，特别是现代物流的发展，将进一步加快世界各大港口的信息化建设速

度和提高信息化的应用水平。信息化建设目标从最初的为码头、船代公司、仓储和货主（货代）之间传递报文、共享信息，逐步扩展到海关、海事、检验检疫等政府部门和银行等金融部门；最终把物流园区、商贸企业、铁路、公路和航空全部纳入系统的服务对象[6]。港口信息化与电子商务的发展目标如图 1-1 所示。

图 1-1　港口信息化与电子商务发展趋势

资料来源：《天津港“十一五”信息化建设发展规划研究》[6]

1.2.2　海运信息应用现状及存在问题

进入 21 世纪以来，随着信息化建设的不断推进，世界各大港口先后建立了口岸信息共享平台，利用互联网发布船舶动态信息，提供各种信息查询服务，加强了相关行业的协作能力，提升了竞争能力。比如，台湾的 5 个主要港口高雄、基隆、台中、花莲和苏澳都各自建立了船舶动态信息预报系统[8-12]，并通过互联网实时发布这些信息，用于指导业务部门调度安排以及为世界各地用户提供查询服务。图 1-2 为台中港务局提供的船舶动态查询界面和显示的查询结果。

图 1-2 台中港务局的船舶动态查询

资料参考：《台中港务局全球资讯网》[10]

通过浏览各个港务网站发布的船舶动态、船期预告和统计报表，我们可以从中获取大量的海运基础信息，通过清理和汇总，加以研究分析可以形成有价值的资料。以台湾港口为例，经过对其 5 个港务局网站的调查整理，我们发现与船舶信息有关的网页共有 33 个，各港口发布的网页信息如表格 1-1。

表格 1-1 台湾港务局船运信息发布情况表

港口名称	网页数量	动态信息网页名称
高雄港	14	最近 24 小时船舶实际进港时间、预定进港船舶、预定进港油轮、船席指泊、公用码头船席调配预排次序、港区船席现况资料表、最近 24 小时船舶实际出港时间、预定出港船舶、船席现况、进港预报次序、最新进出港移泊船期、预定进出港外籍渔船、移锚预报、移泊预报
花莲港	4	表单查询结果、船舶动态、进港次序、出港次序
基隆港	9	船舶动态资讯(30 天)查询、当日船席指派资料查询、船席现况查询、报到、出港、在港动态、进港、优先货柜船船期表、引水人服勤状况、起重机器使用状况
台中港	4	台中港船舶动态表查询、预报进港船舶、在港船舶、预报出港船舶
苏澳港	2	苏澳港船舶动态查询系統、船舶动态

资料来源：《台湾五大港务局网站》[8-12]

通过浏览港务局网站发布的网页信息，可以比较准确地掌握船舶的进港或离港计划，了解港口船舶的数量情况， 为不同部门、不同用户及早安排工作提供了有力帮助。很明显，这些信息是为物流作业提供服务的，而不是为分析统计服务的。要利用这些细节信息进行分析统计，就必须收集全部的记录，经过适当转换和整理后才能实现。比如，我们要查询“某港口在某段时间内哪几个国家的船只来往最多”，就必须将这段时间内所有船舶的进出港记录记录下来，按照船舶所属国家进行分组统计，将统计结果进行排序后方能得出结论。显然，对海运信息的分析和研究绝大部分是针对数据统计来进行的，而能够进行统计的前提是将船舶进出港及相关信息完整地采集下来。

然而，依靠人工浏览查找所需海运信息，利用传统的下载方式来完成信息的收集和整理工作，即使只有几个网页，其工作量之大和困难程度都是难以想象的。具体操作时，需要对相关网页频繁刷新，从大量的发布数据中人工判断是否有新的船舶信息发布，并人工下载保存新数据，再进行后续的数据统计和比对。这种全人工干预的工作模式存在诸多问题：

1. **工作劳动强度大**：需要人工周期性的刷新网页、比对内容、数据下载以及后期的数据统计分析；
2. **网页访问困难**：网站访问经常受限，常常无法正常显示，需要人工不断刷新；并且许多港务局是以分页的形式发布海运数据，即同一份数据分布在多个（最多达到 20 个以上）网页中，需要人工访问每一个网页，如果某一个网页无法访问，将导致数据获取不完整；
3. **人工判断更新困难**：由于数据量大，分布零散，并且更新数据并不是处于固定位置，很可能是在网页中的某一个位置发生了更新，如果以分页形式发布数据，通过人工根本无法判断；
4. **无法及时处理更新数据**：港务局平均每天合计更新的海运数据达到 2000 条左右，人工无法及时处理；
5. **无法掌握海运数据的发布规律**：港务局对某一艘船舶进、出港时间的预报是多批次的，采用人工方式，不可能实时监控海运数据的变化，无法掌握其数据的发布规律；
6. **数据整理困难**：尽管海运数据都是以网页表格的形式发布，但是各个港务局所发布的网页表格格式各不相同，就是同一个港务局内部的格式也不